AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Łańcuchy i zbiory łańcuchów

Teoria automatów i języków formalnych

Dr inż. Janusz Majewski
Katedra Informatyki

Symbol, alfabet, łańcuch

Symbol

Symbol jest to pojęcie niedefiniowane (synonimy: znak, litera)

Alfabet

Alfabet jest to niepusty, skończony zbiór symboli, np.:

$\Sigma_1 = \{a, b, c, ..., x, y, z\}$ - alfabet łaciński

$\Sigma_2 = \{0, 1\}$ - alfabet binarny

Niech Σ będzie alfabetem

$\Sigma \neq \emptyset$; $\#\Sigma < \infty$

Łańcuch

Łańcuch (synonim słowo) to skończony ciąg zestawionych razem symboli z alfabetu.

Rekurencyjna definicja łańcucha nad alfabetem Σ:

1) ε jest łańcuchem nad Σ (ε - łańcuch pusty, nie zawierający żadnego znaku),
2) jeśli x jest łańcuchem nad Σ i $a \in \Sigma$ to xa jest łańcuchem nad Σ,
3) nic innego nie jest łańcuchem poza tym, co wynika z punktów (1) i (2).

Złożenie łańcuchów

Konkatenacja (złożenie) łańcuchów

Jeśli x i y są łańcuchami

$x = x_1x_2...x_m$

$y = y_1y_2...y_n$

to

$xy = x_1x_2...x_my_1y_2...y_n$

jest nazywane konkatenacją (złożeniem) łańcuchów x i y.

Przykład:

x = ab

y = cde

xy = abcde

Dla dowolnego łańcucha x zachodzi

$x\varepsilon = \varepsilon x = x$

Konkatenacja łańcuchów jest:

- nieprzemienna (na ogół $xy \neq yx$)
- łączna ($xyz = (xy)z = x(yz)$)
- posiada element neutralny ε będący łańcuchem pustym ($x\varepsilon = \varepsilon x = x$)

Rewers łańcucha

Rewersem (odbiciem zwierciadlanym) x^R łańcucha $x = x_1x_2...x_n$ nazywamy łańcuch $x^R = x_nx_{n-1}...x_1$, w którym symbole zapisane są w odwrotnym porządku, przy czym $\varepsilon^R = \varepsilon$.

Definicja rekurencyjna odbicia zwierciadlanego:

1) $\varepsilon^R = \varepsilon$,
2) jeśli x jest łańcuchem nad alfabetem Σ, zaś a jest symbolem $(a \in \Sigma)$, to $(ax)^R = x^Ra$,
3) nic innego nie jest odbiciem zwierciadlanym poza tym, co wynika z punktów (1) i (2).

Palindrom

Palindromem nazywamy łańcuch, który czytany od przodu oraz czytany wspak brzmi tak samo, czyli $x^R = x$

Definicja rekurencyjna palindromu:

1) ε jest palindromem,
2) jeśli *a* jest symbolem *(a ∈ Σ)*, to łańcuch zbudowany z pojedynczego symbolu *a* jest palindromem,
3) jeśli *x* jeśli palindromem, zaś *a* jest symbolem *(a ∈ Σ)*, to łańcuch *axa* jest palindromem,
4) nic innego nie jest palindromem poza tym, co wynika z punktów (1), (2) i (3).

Przykłady palindromów: oko, kajak, kobyłamamałybok.

Notacja potęgowa, długość łańcucha

Notacja „potęgowa"

Niech *a* będzie symbolem. Niech *x* będzie łańcuchem. Wtedy łańcuchy złożone z wielokrotnych powtórzeń symbolu *a* lub łańcucha *x* oznaczamy w uproszczeniu:

$a^0 = ε$	$x^0 = ε$
$a^1 = a$	$x^1 = x$
$a^2 = aa$	$x^2 = xx$
$a^3 = aaa$, ... , itd.	$x^3 = xxx$, ..., itd.

Długość łańcucha

Długość $|x|$ łańcucha *x* to liczba symboli tworzących ten łańcuch.

Definicja rekurencyjna długości łańcucha:

1) $|ε| = 0$,
2) jeśli *x* jeśli łańcuchem, zaś *a* jest symbolem *(a ∈ Σ)*, to $|ax| = 1 + |x|$.

Złożenie zbiorów łańcuchów (1)

Niech U_1 i U_2 będą zbiorami łańcuchów nad alfabetami odpowiednio Σ_1 i Σ_2. Złożeniem U_1U_2 tych zbiorów jest zbiór zawierający łańcuchy postaci x_1x_2, gdzie $x_1 \in U_1$, zaś $x_2 \in U_2$.

$U_1U_2 = \{ x_1x_2 \mid x_1 \in U_1, x_2 \in U_2 \}$

Na ogół wygodnie jest utożsamić alfabet ze zbiorem łańcuchów o długości równej jeden każdy, czyli zawsze, gdy nie będzie to budzić wątpliwości, będziemy tak samo oznaczać alfabet i zbiór łańcuchów jednosymbolowych, np.:

$\Sigma = \{ a, b, c \}$ – alfabet, zbiór symboli

$\Sigma = \{ a, b, c \}$ – zbiór łańcuchów jednoliterowych

Złożenie zbiorów łańcuchów jest:

- nieprzemienne (na ogół $U_1U_2 \neq U_2U_1$)
- łączne ($U_1U_2U_3 = (U_1U_2)U_3 = U_1(U_2U_3)$)
- posiada element neutralny $\{\varepsilon\}$ będący zbiorem jednoelementowym, którego jedynym elementem jest łańcuch pusty ($U\{\varepsilon\} = \{\varepsilon\}U = U$)

Złożenie zbiorów łańcuchów (2)

<u>Przykład</u> składania zbiorów łańcuchów:

$\Sigma_1 = \{a, b\}$

$\Sigma_2 = \{a, b, n, r \}$

$U_1 = \{ a, ba \}$

$U_2 = \{ rnaba, rab \}$

$U_1U_2 = \{ arnaba, barnaba, arab, barab \}$

$U_2U_1 = \{ rnabaa, rnababa, raba, rabba \}$

Zbiór słownikowy (1)

Zbiór wszystkich łańcuchów nad alfabetem

Zbiór wszystkich łańcuchów nad alfabetem Σ oznaczamy Σ^*. Zbiór wszystkich niepustych łańcuchów nad alfabetem Σ oznaczamy Σ^+.

Przykład:

$\Sigma = \{0, 1\}$

$\Sigma^* = \{ \varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, ... \}$

$\Sigma^+ = \{ 0, 1, 00, 01, 10, 11, 000, 001, ... \}$

Zbiór słownikowy (2)

Formalnie stosując notację „potęgową" i wykorzystując definicję złożenia zbiorów łańcuchów mamy dla alfabetu Σ definicję rekurencyjną:

$\Sigma^0 = \{\varepsilon\}$

$\Sigma^1 = \Sigma = \{ x \mid x$ jest słowem nad Σ, $|x| = 1 \}$

$\Sigma^2 = \Sigma^1\Sigma$

....

$\Sigma^n = \Sigma^{n-1}\Sigma = \{ x \mid x$ jest słowem nad Σ, $|x| = n \}$

$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$

$\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$

$\Sigma^+ = \Sigma\, \Sigma^*$

Uporządkowanie zbioru słownikowego (1)

Niech $\leq$ będzie relacją liniowego porządku na zbiorze (alfabecie) Σ.

Przykład: $\Sigma=\{a,b,c,d\}$. Uporządkowanie „alfabetyczne": a≺b; a≺c; a≺d; b≺c; b≺d; c≺d

Zdefiniujemy relację $\leq_S$ określoną na Σ^*. Powiemy, że $x = a_1a_2...a_m$ jest w relacji $\prec_S$ z $y = b_1b_2...b_n$ ($x \prec_S y$; $x,y \in \Sigma^*$; $a_i,b_j \in \Sigma$ dla $1 \leq i \leq m$, $1 \leq j \leq n$), gdy spełniony jest jeden z dwóch poniższych warunków:

- $m < n$,
- $m = n$ oraz $a_i \prec b_i$ dla pewnego $i \leq m=n$ oraz $a_j = b_j$ dla wszystkich $1 \leq j < i$.

Relacja $\prec_S$ jest quasi-porządkiem, który w zbiorze Σ^* definiuje porządek $\leq_S$ nazywamy porządkiem standardowym. $(\Sigma^*, \leq_S)$ jest zbiorem dobrze uporządkowanym.

Przykład: $\Sigma=\{a,b\}$. Porządek alfabetu: a≺b. Zbiór słownikowy Σ^* w porządku standardowym:

$\Sigma^*=\{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, bab\ldots\}$

Uporządkowanie zbioru słownikowego (2)

Niech $\leq$ będzie relacją liniowego porządku na zbiorze (alfabecie) Σ.

Zdefiniujemy relację $\leq_L$ określoną na Σ^*. Powiemy, że $x = a_1a_2...a_m$ jest w relacji $\prec_L$ z $y = b_1b_2...b_n$ ($x \prec_L y$; $x,y \in \Sigma^*$; $a_i,b_j \in \Sigma$ dla $1 \leq i \leq m$, $1 \leq j \leq n$, $k = \min(m,n)$), gdy spełniony jest jeden z dwóch poniższych warunków:

(1) $a_i \prec_L b_i$ dla pewnego $i \leq k$ oraz $a_j = b_j$ dla wszystkich $1 \leq j < i$

(2) $m < n$ oraz $a_i = b_i$ dla wszystkich $1 \leq i \leq m=k$.

Relacja $\prec_L$ jest quasi-porządkiem, który w zbiorze Σ^* definiuje porządek $\leq_L$, który nazywamy porządkiem leksykograficznym. $(\Sigma^*, \leq_L)$ jest zbiorem liniowo uporządkowanym. Nie jest on jednak zbiorem dobrze uporządkowanym.

Uporządkowanie zbioru słownikowego (3)

Przykład: $\Sigma=\{a,b\}$. Porządek alfabetu: $a \prec b$. Kilka pierwszych elementów zbioru słownikowego Σ^* w porządku leksykograficznym:

$\Sigma^*=\{\varepsilon, a, aa, aaa, aaaa, aaaaa, \ldots\}$

Dlaczego uporządkowanie leksykograficzne nie jest porządkiem dobrym? Weźmy nieskończony podzbiór zbioru Σ^*: {b, ab, aab, aaab, aaaab, …}. Elementy tego podzbioru są uporządkowane malejąco:

$b \succ_L ab \succ_L aab \succ_L aaab \succ_L aaaab \succ_L \ldots$

Oczywiście podzbiór ten nie ma elementu najmniejszego, uporządkowanie leksykograficzne nie jest więc porządkiem dobrym. Porządek leksykograficzny w nieskończonym zbiorze Σ^* jest bardzo skomplikowany i trudno go sobie wyobrazić.

Uporządkowanie zbioru słownikowego (4)

Porządek leksykograficzny jest jednak powszechnie wykorzystywany do skończonych zbiorów łańcuchów, np. do określania kolejności słów w encyklopediach, słownikach, leksykonach. Wówczas quasi-porządek $\prec$ jest powszechnie przyjętym uporządkowaniem liter w alfabecie pewnego języka naturalnego.

Ponieważ dla zbioru słownikowego Σ^* można określić porządek liniowy (np. leksykograficzny) lub porządek dobry (np. standardowy), można więc wszystkie elementy zbioru słownikowego ułożyć w ciąg i ponumerować. Świadczy to o równoliczności zbioru słownikowego ze zbiorem liczb naturalnych.

Uporządkowanie zbioru słownikowego (5)

Numeracja Gödla: Można ponumerować elementy zbioru słownikowego:

- numerujemy elementy alfabetu $\Sigma=\{a_1, a_2, \ldots, a_n\}$
- niech $p_1 p_2 p_3 p_4 \ldots$ będzie ciągiem rosnącym liczb pierwszych, np. $2, 3, 5, 7, \ldots$
- określamy funkcję $num(x)$ dla $x \in \Sigma^*$

$$num(\varepsilon) = 0$$

$$num(a_{i_1} a_{i_2} \ldots a_{i_k}) = \prod_{j=1}^{k} p_j^{\,i_j}$$

Można pokazać, że odwzorowanie

$num: \Sigma^* \mapsto \boldsymbol{N}$

jest wzajemnie jednoznaczne (funkcja $num(x)$ jest różnowartościowa).

Przykład:

$\Sigma=\{a,b\}$, numerujemy litery => $\Sigma=\{a_1,a_2\}$

określamy rosnący ciąg liczb pierwszych: $p_1, p_2, p_3, \ldots$ jako $2, 3, 5, 7, \ldots$

analizujemy słowo $x = abaa \in \Sigma^*$, $x = a_1 a_2 a_1 a_1$,

$num(x) = p_1^1 \cdot p_2^2 \cdot p_3^1 \cdot p_4^1 = 2^1 \cdot 3^2 \cdot 5^1 \cdot 7^1 = 2 \cdot 9 \cdot 5 \cdot 7 = 630$